Jurnal CULTURE (Culture, Language, and Literature Review), 12(2), November 2025, 123-138

e-ISSN 2775-4618, p-ISSN 2355-8660, bit.ly/jurnalculture / culture@unaki.ac.id

Item analysis in a psycholinguistics course based on classical test theory using ITEMAN 4.0.2

Siska Adinda Prabowo Putri[1)], Lucy Hariadi[2)], Amin Khudlori[3)]
[1]Faculty of Psychology, AKI University
e-mail: sisca.adinda@unaki.ac.id
[2]Faculty of Psychology, AKI University
e-mail: lucy.hariadi@unaki.ac.id
[3] Faculty of Language & Culture, AKI University
e-mail: amin.khudlori@unaki.ac.id

Abstract

This study aims to analyze the quality of test items in a psycholinguistics course using Classical Test Theory. The data used consisted of 30 students' responses to 20 multiple-choice test items, which were analyzed using several indicators: difficulty level, discriminatory power, item-total correlation, and instrument reliability. The results showed that most items were in the moderate difficulty category (p = 0.46–0.56), with one item categorized as easy (p = 0.73) and two items as difficult (p = 0.26). The discriminatory power of the majority of items was in the very good category (87.5%–100%), while three items showed lower discriminatory power and required revision. The item-total correlation was generally very high (r = 0.88–0.99), indicating consistency among items, but several items with lower correlations (r < 0.70) suggested possible wording inaccuracies or content inconsistencies. The test's reliability reached 0.99, indicating very high internal consistency, although this value was influenced by the quite extreme response patterns between the upper and lower groups. Overall, the test instrument was considered good, but several items needed revision, particularly in terms of distractors, difficulty level, and item functionality, to ensure more accurate and representative learning evaluations.

Keywords: Classical test theory, Item analysis, ITEMAN 4.0.2, Psycholinguistics, Reliability

1. Introduction

Learning evaluation is an important component of the educational process because it provides objective information on student achievement and the effectiveness of teaching. One of the most widely used evaluation instruments in higher education is the objective test, especially the multiple-choice form. Although it is easy to manage and assess, test quality is not determined solely by the number of items; it depends on the quality of each item. Therefore, item analysis is an essential step to ensure that each question can measure abilities accurately, fairly, and consistently (Farida & Musyarofah, 2021)

Item analysis plays a fundamental role in learning evaluation because it determines the extent to which a test instrument provides an accurate picture of student abilities. In language education, especially Psycholinguistics courses, the quality of items not only measures theoretical comprehension, but also cognitive abilities such as perception, language processing,

e-ISSN 2775-4618, p-ISSN 2355-8660, bit.ly/jurnalculture / culture@unaki.ac.id

working memory, and mental representation (Traxler, 2023). Therefore, assessment instruments must be systematically designed and analyzed to ensure valid and reliable results.

In line with the development of assessment technology, the use of software such as ITEMAN 4.0 is increasingly recommended. ITEMAN 4.0.2 presents an analysis based on Classical Test Theory (CTT), which includes difficulty index, differentiating power, trick effectiveness, test reliability, and distribution of student scores (Thompson, 2022). The main advantage of ITEMAN lies in the ease of interpreting its outputs, which is very helpful for lecturers in identifying good question items, those that need revision, or those that must be discarded. The program also minimizes manual analysis errors that often occur in large-scale data processing (Hrich et al., 2024).

In the context of the Faculty of Language and Culture of UNAKI, the Psycholinguistics course requires students to understand the mental processes underlying language use, including phonological processing, semantics, syntax, speech comprehension, and the relationship between language and cognition. Therefore, assessment instruments must be able to measure students' abilities in these various domains validly. However, based on initial observations, there are still items that have unbalanced difficulties, have low discriminating power, and use tricks that do not function optimally. This condition can cause the evaluation results not to reflect their true abilities (Brown & Abeywickrama, 2021; Alderson, 2020).

Conducting item analysis using ITEMAN 4.0.2 is a strategic step to improve the quality of Psycholinguistics learning evaluation. Recent researches also show that the application of digital analytics can improve evaluation quality, enhance question banks, and support continuous assessment in higher education (Hrich et al., 2024; Hartati & Yogi, 2019). Thus, this research not only aims to evaluate the quality of the questions used but also contributes to the development of technology-based assessments at UNAKI. In addition, modern software-assisted grain analysis approaches are in line with the needs of colleges to raise academic assessment standards. With the empirical data generated from ITEMAN 4.0.2, Psycholinguistics lecturers can improve instruments more systematically, increase the validity of test construction, and ensure that the questions used are unbiased and in accordance with the learning outcomes of the study program.

Some contemporary studies emphasize that item analysis is not just an additional procedure but an integral part of the instrument's evaluation and development cycle. For example, recent research shows that even if a test is declared globally reliable (high α/KR-20), many individual items have low differentiating power or are ineffective, thereby reducing the test's overall quality (Putri, et al, 2024). Several previous studies have also strengthened that CTT is relevant and popular as a grain analysis approach, such as Ohiri & Okoye (2024), Rohmatdi (2024), Resi (2023), Subhaktiyasa (2024), Liu & Maydeu-Olivares (2024), who reviewed the application of CTT and ITEMAN in the development and analysis of grains, question difficulty methods, discrimination, and reliability as standard test procedures.

Based on this urgency, this study aims to analyze the quality of question items in the Psycholinguistics course at the Faculty of Languages and Culture of UNAKI using the ITEMAN 4.0.2 Program, identify the level of difficulty of the question items, assess the differentiating power of the question items to determine the ability of the question items in distinguishing high and low ability students, analyze the effectiveness of the deception on multiple-choice items, determine the overall reliability of the test based on The results of the ITEMAN 4.0.2 analysis are also the basis for improvements in the preparation of question items so that the evaluation is more valid, reliable, and representative of student competence.

e-ISSN 2775-4618, p-ISSN 2355-8660, bit.ly/jurnalculture / culture@unaki.ac.id

2. Theoretical Framework

2.1 Classical Test Theory (CTT)

Classical Test Theory (CTT) is the most widely used measurement approach in educational and psychological research, especially in the development of achievement test instruments. Within this theory, test quality is evaluated by the relationship between a participant's apparent score and his or her true score. CTT departs from the basic assumption that a person's test score does not fully reflect true ability due to error, whether arising from the participant, the environment, or the instrument. Miller and Lovler (2020) explain that CTT is based on the main formula X = T + E, where X is the observed score, T is the true score, and E is the error component. Errors in CTT are assumed to be random and unsystematic, so they can be minimized but not eliminated.

Further, CTT assumes that measurement errors do not correlate with participants' true abilities. In this context, high-ability participants are no more likely to make mistakes than low-ability participants. This assumption makes CTT relatively simple and easy to use to assess the quality of test instruments. According to DeVellis and Thorpe (2021), CTT assumptions allow the evaluation process of instruments to be carried out with basic statistical techniques so that they can be applied in various types of research, including small-scale classroom tests.

In practice, evaluating the quality of test items using CTT usually involves four main components: difficulty index, discrimination index, item-total correlation, and instrument reliability. The level of difficulty is the proportion of participants who correctly answer the item. Items that are too easy or too difficult often do not measure the variation in participants' abilities optimally. Allen and Yen (2022) emphasized that the ideal difficulty level lies in the medium category because it provides the most informative information about participants' abilities.

In addition, CTT assesses the differentiating power, which indicates an item's ability to distinguish between high- and low-ability participants. Good grains must have a high, positive differentiating power. Low or negative differentiating power indicates that the item does not function as intended, as it does not provide valid information about participants' abilities. In some cases, low differentiation indicates that a high-ability participant fails to answer an item. In contrast, a low-ability participant answers it correctly—a strong sign that the item is editorial or substantive. Grain quality can also be assessed through point-biserial correlation. A high correlation indicates that the item is consistent with the overall test. In contrast, low correlation usually indicates that the item measures aspects that differ from the overall instrument or contains distractions, such as question ambiguity. Therefore, items with low correlation are often recommended for revision or removal.

Another most important component of CTT is the test's reliability, which is the extent to which the instrument provides consistent results when repeated or used with a similar group of participants. For multiple-choice objective tests, reliability calculations often use the Kuder-Richardson formula of 20 (KR-20), which is indeed designed for dichotomous items. High reliability indicates that the instrument measures consistently, while low reliability indicates inconsistent measurement in participants' abilities. Kline (2020) emphasises that reliability is a necessary condition for validity; without adequate reliability, test results cannot guarantee that the score truly reflects participants' abilities.

Thus, Classical Test Theory has an important position and remains relevant in the development of educational instruments. Advantages in simplicity of analysis, ease of interpretation, and suitability for small scales make CTT a top choice for educators and researchers. However, its use must be accompanied by an understanding of its limitations to avoid misinterpretation of the item evaluation results. The combination of CTT and modern

e-ISSN 2775-4618, p-ISSN 2355-8660, bit.ly/jurnalculture / culture@unaki.ac.id

measurement methods will provide a stronger foundation in the development of performance tests that are valid, reliable, and capable of providing an accurate picture of participant competencies.

2.2 Test Difficulty Level

The difficulty level is calculated to find out the proportion of participants who can answer the question correctly, using the formula:

$$p = \frac{B}{N}$$

Where:
B = number of participants who answered yes,
N = total number of participants.

Category interpretations: Easy ($p \geq 0.70$), Medium ($0.30 \leq p < 0.70$), Difficult ($p < 0.30$). In this study, most items are in the medium category, with one easy item and two difficult items.

2.3 Test Discriminating Power

Differentiating power is used to determine the ability of the grain to distinguish between high- and low-ability participants. The upper and lower groups each accounted for 27% of the total participants (the top 8 students and the bottom 8).

Calculation formula:

$$D = \frac{B_A - B_B}{N/2}$$

Where:
B_A = the correct number in the top group,
B_B = the correct amount in the bottom group.

Interpretation: *Excellent* ($D \geq 0.40$), *Good* (0.30–0.39), *Adequate* (0.20–0.29), *Poor* ($D <$ 0.20). The results showed that most grains had very good differentiating power, while some grains had lower differentiating power and needed revision.

2.4 Effectiveness of Distractors

The distraction was analyzed to determine whether each trick option was working properly, i.e., whether it was chosen by participants who had not mastered the material. An effective distraction when:

1. Selected by at least **5% of participants**,
2. It is not chosen more by the upper group than the lower group,
3. No answer option is selected at all (the distressor is dead)

e-ISSN 2775-4618, p-ISSN 2355-8660, bit.ly/jurnalculture / culture@unaki.ac.id

2.5 Instrument reliability

Reliability was calculated using the Kuder–Richardson Formula 20 (KR-20) because the instrument consisted of multiple-choice questions with dichotomous scores (1–0).

KR-20 Formula:

$$KR20 = \frac{k}{k-1}(1 - \frac{\sum pq}{\sigma^2})$$

Where:
k = number of grains,
p = true proportion per item,
q = 1 - p,
σ² = total score variance.

The test's reliability in this study was 0.99, indicating very high internal consistency.

4. Research Method

This study uses a cross-sectional, quantitative approach, applying the Classical Test Theory (CTT) framework. This approach was chosen because it can provide statistical information about the characteristics of each question item by calculating the level of difficulty, differentiation, distractor effectiveness, and instrument reliability. The research subjects were 30 students enrolled in the Psycholinguistics course at the Faculty of Languages and Letters. The instruments analyzed were in the form of 20 multiple-choice questions with four answer options (A–D). The design of the question grid is shown in Table 1.

Table 1. Psycholinguistics Question Grid Design

Yes	Learning Outcome (LO)	Material/ Competencies	Cognitive Level	Number of Questions	Question Number
1	LO-1: Understanding basic concepts	Definition, scope, brain-language relationship	C1–C2	4	1, 2, 3, 4
2	LO-2: Analyzing language processing	Speech production, perception, mental lexicon	C2–C4	5	5, 6, 7, 8, 9
3	LO-3: Understanding first language acquisition	L1 stages, models, child language	C1–C3	4	10, 11, 12, 13
4	LO-4: Understanding second language acquisition & bilingualism	L2 theories, bilingualism, code-switching	C2–C4	4	14, 15, 16, 17
5	LO-5: Identifying language disorders	Dyslexia, speech disorders	C1–C3	3	18, 19, 20

e-ISSN 2775-4618, p-ISSN 2355-8660, bit.ly/jurnalculture / culture@unaki.ac.id

Data were collected from students' written exam scores. Each participant's response is given a score **of 1** if true and **0** if false. The data is then entered into the item analysis sheet for further processing using the CTT formula. Data analysis using Anates 4.0.2 software

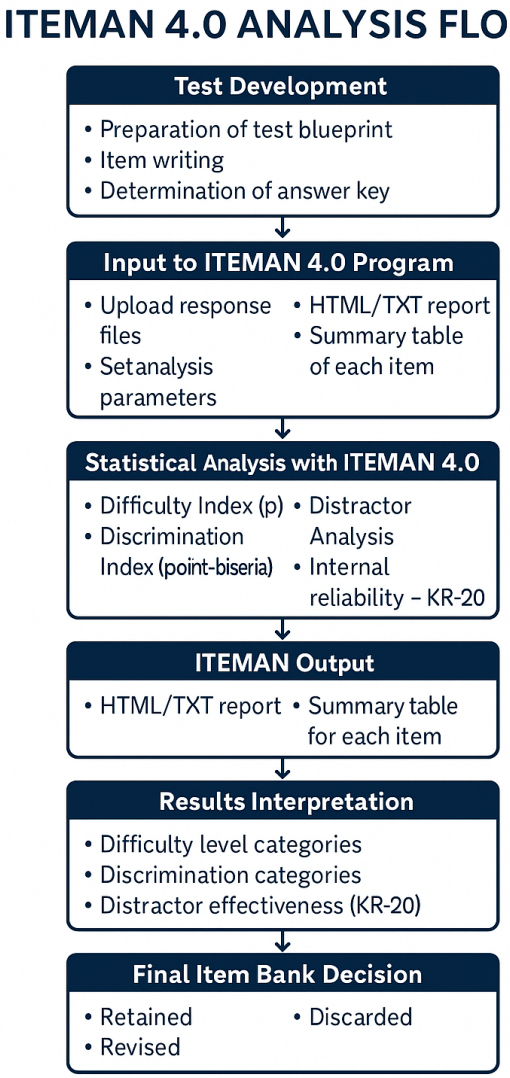


Figure 1. Iteman 4.0 Analysis Flowchart

4. Findings and Discussions

4.1 Distribution of Respondent Answers

In Figure 2, it can be seen that the respondents in this study were 30 people, with a total of 20 questions. From Figure.1, it can also be seen that the pattern of respondents' answers varies across all item distributions.

e-ISSN 2775-4618, p-ISSN 2355-8660, bit.ly/jurnalculture / culture@unaki.ac.id

Jumlah Subyek 30 Jumlah Butir Soal 20 Jumlah Pilihan Jawaban 4 Tips: Gunakan tombol ENTER untuk pindah antar kolom

No.Urut	Kode/Nama Subyek	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
KUNCI->	KUNCI ->	B	C	B	B	B	C	B	C	B	A	B	D	C	B	C	B	C	B	C	C
1	NOVA ANDI	C	C	A	A	A	A	D	B	C	D	A	C	D	C	A	C	A	A	D	D
2	ELLEN DIAN KRISTIAN	C	C	A	C	A	B	D	B	C	D	A	C	D	A	A	C	A	D	D	D
3	VANIA HENVER KURNIAWAN	C	D	A	D	A	A	D	B	C	D	A	C	D	D	A	C	B	C	D	D
4	KIKI NABILA PUTRI SAGITA	C	A	A	A	C	B	D	B	C	D	A	C	D	A	A	C	D	A	D	D
5	PUTRI SINTIYA	C	C	A	A	A	A	D	B	C	D	A	C	D	C	A	C	A	C	D	D
6	SELLY APRILIA	A	D	A	C	A	B	D	B	C	D	A	C	D	D	A	C	B	D	D	D
7	FIDEL DAFA AKBAR	A	C	A	D	C	A	D	B	C	D	A	C	D	A	A	C	D	A	D	D
8	SINTYA PUTRI RAHAYU	A	B	A	A	D	B	D	B	C	D	A	C	D	D	A	C	A	C	D	D
9	FARICHA AZ-ZAHRA DWI FEBRIYANI	B	C	A	C	C	A	D	B	C	D	A	C	D	A	A	C	B	D	D	D
10	NUR MUTIA KUMALASARI	B	A	A	D	A	B	D	B	C	D	A	C	D	C	A	C	D	A	D	D
11	SITI NURUUS SAADAH	A	C	A	A	D	A	D	B	C	D	A	C	D	C	A	C	A	C	D	D
12	AYU NOVITASARI	A	D	A	C	D	B	D	B	C	D	A	C	D	A	A	C	B	D	D	D
13	DYAS GALUH JATININGSIH	A	C	A	D	C	A	D	B	C	D	A	C	D	D	A	C	D	A	D	D
14	REVHA AZHIRA ZHAHWA	A	B	A	A	C	B	D	B	C	D	A	C	D	D	A	C	A	C	D	D
15	SHINTA DWI RAHAYU	A	D	A	C	D	A	D	B	C	D	A	C	D	D	A	C	B	D	D	D
16	SITI SOLEKHAH	B	C	B	B	B	C	A	A	B	A	B	B	C	B	C	B	C	B	C	C
17	ABEL HIDAYAT	B	C	B	B	B	C	B	C	B	A	B	D	C	B	C	B	C	B	C	C
18	NANDA CAHYA BINTANG	B	C	B	B	B	C	B	C	B	A	B	D	C	B	C	B	C	B	C	C

Tambah Subyek Sisipkan Subyek Hapus Subyek Hapus Butir Soal

Figure 2. Respondents' Answers

4.2 Test difficulty level

The difficulty level (p) is the proportion of test takers who answered a question correctly. This value indicates how easy or difficult an item is for a particular group of participants. The more participants answered correctly, the easier it would be; On the other hand, the fewer who answered yes, the more difficult the details became. According to Azwar (2017), the interpretation of the test difficulty value is divided into three, namely if the p value is < 0.30, then the problem is classified as difficult, if the value is $0.30 \leq p \leq 0.70$, then it is classified as moderate, and if the p value is > 0.70, then it is relatively easy. Azwar (2017) emphasized that items that are too easy or too difficult are not informative in distinguishing participants based on ability.

e-ISSN 2775-4618, p-ISSN 2355-8660, bit.ly/jurnalculture / culture@unaki.ac.id

Tingkat Kesukaran Kembali Ke Menu Utama Ce

Jml Subyek= 30 Butir Soal = 20

No Butir	Jml Betul	Tkt. Kesukaran(%)	Tafsiran
1	17	56.67	Sedang
2	22	73.33	Mudah
3	15	50.00	Sedang
4	14	46.67	Sedang
5	15	50.00	Sedang
6	15	50.00	Sedang
7	8	26.67	Sukar
8	14	46.67	Sedang
9	15	50.00	Sedang
10	15	50.00	Sedang
11	15	50.00	Sedang
12	14	46.67	Sedang
13	15	50.00	Sedang
14	8	26.67	Sukar
15	11	36.67	Sedang
16	15	50.00	Sedang
17	15	50.00	Sedang
18	15	50.00	Sedang
19	15	50.00	Sedang
20	11	36.67	Sedang

Figure 3. Test Difficulty Level

The best items are those in the medium category, as they can yield a range of scores and help distinguish between high- and low-ability participants. In the 20 questions of the Psychogisticistics Test, there were 17 items (1, 3, 4, 5, 6, 8, 9, 10, 11, 12, 13, 15, 16, 17, 18, 19, 20) which had a moderate category test difficulty level (p = 0.36 – 0.56), one question item (number 2) was included in the easy category (p = 0.73), two questions (7 and 14) were included in the difficult category (p = 0.26). The test difficulty results can also be seen in Figure 3.

4.3 Test the discriminative power

Discriminating power (symbolized by D) is the ability of a question item to distinguish between participants who have high and low ability. The higher the power of discrimination, the better the item is at identifying who really controls the material. In general, the discriminating power is calculated by comparing:

Proportion of high-ability participants (top group) who answered yes

Proportion of low-ability participants (lower group) who answered yes

According to Azwar (2017) and Ebel & Frisbie (1991), the test discrimination score categories are shown in Table 1.

e-ISSN 2775-4618, p-ISSN 2355-8660, bit.ly/jurnalculture / culture@unaki.ac.id

Table 1. Categories of Discriminatory Power Values (D)

D Value	Category	Interpretation
≥ 0.40	**Excellent**	Very strong differentiating ability
0.30 – 0.39	Good	Quite differentiating
0.20 – 0.29	Enough	Need for revision
0.00 – 0.19	Bad	Grains are not able to distinguish
Negative	Very Bad	The lower participants answered correctly → problematic items

The relationship between the ability to discriminate and the level of difficulty of the test can be seen in items that are very easy or very difficult; they usually have low discriminating power because almost all participants are right or wrong. Medium category items most often result in high discrimination. Therefore, the preparation of a good test always involves a balanced combination of difficulty level and discriminating power. In CTT, discriminating power is an important parameter because:

1. Contributes to the empirical validity of the item (Crocker & Algina, 2008)
2. Improves the overall reliability of the test
3. Ensuring the instrument truly measures ability, not just luck
4. Determining which items to keep, revise, or discard
5. Tests with high discriminating power provide more accurate data for assessment and research.

Daya Pembeda Kembali Ke Menu Utama

Jml Subyek= 30 Klp atas/bawah (n) = 8 B

No Butir	Kel. Atas	Kel. Bawah	Beda	Indeks DP (%)
1	8	0	8	100.00
2	8	1	7	87.50
3	8	0	8	100.00
4	8	0	8	100.00
5	8	0	8	100.00
6	8	0	8	100.00
7	7	0	7	87.50
8	8	0	8	100.00
9	8	0	8	100.00
10	8	0	8	100.00
11	8	0	8	100.00
12	8	0	8	100.00
13	8	0	8	100.00
14	4	0	4	50.00
15	8	0	8	100.00
16	8	0	8	100.00
17	8	0	8	100.00
18	8	0	8	100.00
19	8	0	8	100.00
20	8	0	8	100.00

Figure 4. Test Discriminating Power

e-ISSN 2775-4618, p-ISSN 2355-8660, bit.ly/jurnalculture / culture@unaki.ac.id

In Figure 4, it can be seen that of the 20 questions in the Psycholinguistic Test, 19 items have very good Discriminating Power (D = 87.5% - 100%), and 1 item (number 14) has a Discriminating Power of 50%. It can be concluded that the instrument is very effective in distinguishing between the upper and lower groups: the high-achievement group answers almost all items, whereas the low-achievement group fails to answer most. In CTT theory, a differentiating power of ≥ 0.40 is considered good and ≥ 0.70 is considered excellent (Ebel & Frisbie, 1991). So that this instrument discriminates very strong tests

4.4 Distractor effectiveness

A distractor is an answer option other than the correct answer to a multiple-choice question. According to Haladyna (2004), a good deceiver is one chosen proportionally by low-ability participants, not by high-ability participants. Meanwhile, according to Nitko & Brookhart (2011), a deceiver who does not function is chosen by less than 5% of participants, has never been selected, or is chosen more by the upper group than by the lower group. The results of the ITEMAN analysis can be seen in Figure 5, which shows that:

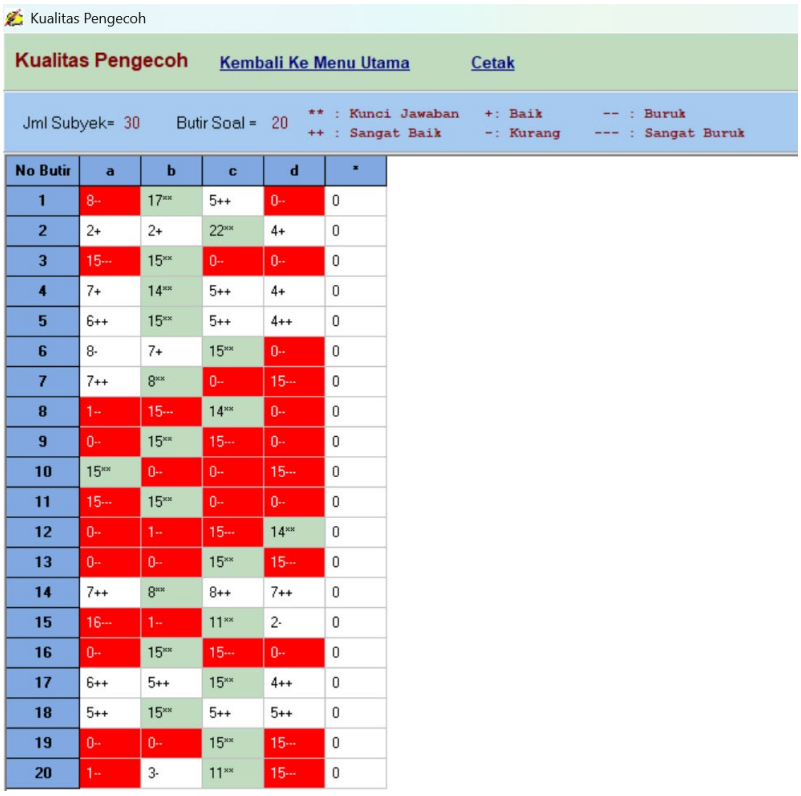
1. Non-functioning divertors (--- or --)
   Point 1: options a (8–) and d (0–)

   Item 3: a (15---), c (0--)
   Items 10, 11, 19: b (0--), c (0--), d (15---)
   Item 12: a (0--), b (1--)
   Meaning:
   a) Participants hardly choose the option
   b) Distracters do not interfere with low-ability participants
   c) Grain quality decreases
   d) Indicates an item is too easy for the upper group
   e) Indications of teaching effect or too familiar questions
2. Wrong Extroverts Selected by the Top Group
   Item 2: option a (+), b (+), d (+)
   Item 4: a (+), c (++), d (+)
   Item 5: a (++), c (++), d (++)
   Meaning:
   a) Low-ability participants are interested in choosing a trickster
   b) The distractor works according to the test theory
   c) Grains have a deceptive balance
   d) Increase differentiating power
3. Wrong Diverters Selected Upper Group (Items 7 & 14)

According to Haladyna (2004) and Nitko & Brookhart (2011), non-functioning tricksters should be revised or replaced, as they reduce psychometric quality. Questions with bad tricksters can make the test more predictive for the high group, but unfair for the low group. It can be concluded that ± 50% of the tricksters are not chosen at all. Some of the distractions in these tests did not work, which is in line with previous findings that low-performing distractions reduce the diagnostic power of multiple-choice tests (Aljabr, 2020). However, there are still items with excellent tricksters. The recommendations for improvement include

e-ISSN 2775-4618, p-ISSN 2355-8660, bit.ly/jurnalculture / culture@unaki.ac.id

that the researcher revise the deceiver that was never selected, add a more plausible deceiver (similar to the answer key), and avoid the deceiver whose answer is too obvious.



Kualitas Pengecoh

Kualitas Pengecoh Kembali Ke Menu Utama Cetak

Jml Subyek= 30 Butir Soal = 20 ** : Kunci Jawaban +: Baik -- : Buruk ++ : Sangat Baik -: Kurang --- : Sangat Buruk

No Butir	a	b	c	d	*
1	8--	17**	5++	0--	0
2	2+	2+	22**	4+	0
3	15---	15**	0--	0--	0
4	7+	14**	5++	4+	0
5	6++	15**	5++	4++	0
6	8-	7+	15**	0--	0
7	7++	8**	0--	15---	0
8	1--	15---	14**	0--	0
9	0--	15**	15---	0--	0
10	15**	0--	0--	15---	0
11	15---	15**	0--	0--	0
12	0--	1--	15---	14**	0
13	0--	0--	15**	15---	0
14	7++	8**	8++	7++	0
15	16---	1--	11**	2-	0
16	0--	15**	15---	0--	0
17	6++	5++	15**	4++	0
18	5++	15**	5++	5++	0
19	0--	0--	15**	15---	0
20	1--	3-	11**	15---	0

Figure 5. Traits of Tricksters

4.5 Total correlation item results

Item–total correlation measures how well an item is related to, or contributes to, the construct measured by the entire test. In CTT, this is used as an indicator of an item's internal validity (Azwar, 2017; Crocker & Algina, 2008). Practically, the higher the correlation, the more consistent the item's answer is with the overall test; a low value indicates an item may be irrelevant, ambiguous, or measure another construct.

e-ISSN 2775-4618, p-ISSN 2355-8660, bit.ly/jurnalculture / culture@unaki.ac.id

Jml Subyek= 30 Butir Soal = 20 Info ten

No Butir	Korelasi	Signifikansi
1	0.882	Sangat Signifikan
2	0.631	Sangat Signifikan
3	0.993	Sangat Signifikan
4	0.931	Sangat Signifikan
5	0.993	Sangat Signifikan
6	0.993	Sangat Signifikan
7	0.669	Sangat Signifikan
8	0.939	Sangat Signifikan
9	0.993	Sangat Signifikan
10	0.993	Sangat Signifikan
11	0.993	Sangat Signifikan
12	0.939	Sangat Signifikan
13	0.993	Sangat Signifikan
14	0.610	Sangat Signifikan
15	0.812	Sangat Signifikan
16	0.993	Sangat Signifikan
17	0.993	Sangat Signifikan
18	0.993	Sangat Signifikan
19	0.993	Sangat Signifikan
20	0.812	Sangat Signifikan

Figure 6. Total Correlation Item Results

In the data seen in Figure 6, all question items have an item-total correlation value of > 0.60, so theoretically all items contribute to the measured construct or can be interpreted to have high internal validity. According to Nunnally & Bernstein (1994), values of r ≥ 0.40 are usually considered adequate/good for item–total correlations; ≥ 0.60 is relatively strong; ≥ 0.80–0.90 is very strong.

4.6 Test reliability results

KR-20 is a reliability formula in Classical Test Theory (CTT) used to measure the internal consistency of a dichotomous instrument (true = 1, false = 0). The interpretation of the KR-20 value is shown in Table 2.

Table 2. Interpretation of Reliability

KR-20	Interpretation
0.90 – 1.00	Very High / Excellent
0.80 – 0.89	Tall
0.70 – 0.79	Enough
0.60 – 0.69	Low
< 0.60	Unreliable

The results in this study showed a KR-20 of 0.99, indicating very high reliability. This instrument is very consistent; the error measurement is very small. All items measure the same ability. In addition, the odd-even correlation of 0.98 from the split-half reliability test is also interpreted as high reliability. Where both odd and even item parts measure the same construct and are very consistent.

e-ISSN 2775-4618, p-ISSN 2355-8660, bit.ly/jurnalculture / culture@unaki.ac.id

Rata2= 9.47 Simpang Baku= 9.08 KorelasiXY= 0.98 Reliabilitas Tes = **0.99**

No.Urut	Kode/Nama Subyek	Skor Ganjil	Skor Genap	Skor Total
1	NOVA ANDI	1	0	1
2	ELLEN DIAN KRISTIAN	1	0	1
3	VANIA HENVER KURNIAWAN	0	0	0
4	KIKI NABILA PUTRI SAGITA	0	0	0
5	PUTRI SINTIYA	1	0	1
6	SELLY APRILIA	0	0	0
7	FIDEL DAFA AKBAR	1	0	1
8	SINTYA PUTRI RAHAYU	0	0	0
9	FARICHA AZ-ZAHRA DWI FEBRIYANI	1	1	2
10	NUR MUTIA KUMALASARI	0	1	1
11	SITI NURUUS SAADAH	1	0	1
12	AYU NOVITASARI	0	0	0
13	DYAS GALUH JATININGSIH	1	0	1
14	REVHA AZHIRA ZHAHWA	0	0	0
15	SHINTA DWI RAHAYU	0	0	0
16	SITI SOLEKHAH	7	9	16
17	ABEL HIDAYAT	9	10	19
18	NANDA CAHYA BINTANG	9	10	19
19	MUHAMMAD ZULFA KAMAL	9	10	19
20	DEWI MELATI SARI	8	9	17
21	SUTRIMAH	9	9	18

Figure 2. Test Reliability Results

4.7 Discussion

The results of the item analysis using Classical Test Theory and the ITEMAN 4.0.2 program indicate that the psycholinguistics test instrument generally demonstrates strong psychometric properties. The majority of the 20 items fall within the moderate difficulty level (p = 0.36–0.56), suggesting that the items were neither too easy nor too difficult for most test takers. According to Allen and Yen (2022), items within the medium difficulty category are preferable because they provide optimal score variation and contribute more effectively to distinguishing students' levels of mastery. This finding aligns with Azwar's (2017) recommendation that items in the medium range provide the most informative assessment.

The discriminating power further reinforces this conclusion. Nineteen items demonstrated very high discrimination (87.5%–100%), indicating strong differentiation between high- and low-performing students. High discrimination values suggest that the items effectively assess student ability (Ebel & Frisbie, 1991; Crocker & Algina, 2008). Only one item showed lower discrimination (50%), meaning revision is needed to align the item more closely with learning objectives or reduce ambiguity. Strong discriminating power also contributes to greater test validity and supports accurate decision-making in assessment contexts (Thompson, 2022).

The item–total correlation results also showed high values for all items (r = 0.88–0.99), indicating internal consistency and strong alignment between individual items and the overall construct measured by the test. Nunnally and Bernstein (1994) emphasize that item–total correlations above 0.60 reflect strong item quality. The high correlation values confirm that each item contributes substantially to the assessment of students' psycholinguistic competence. However, several items with slightly lower correlations ($r < 0.70$) may benefit from refinement to ensure clarity and precision in measuring targeted sub-skills.

The test's reliability coefficient (KR-20 = 0.99) reflects exceptionally high internal consistency. According to Kline (2020), such a high reliability index indicates that measurement error is minimal and the instrument consistently assesses the same construct. While high reliability is desirable, values approaching 1.00 may also indicate item redundancy

e-ISSN 2775-4618, p-ISSN 2355-8660, bit.ly/jurnalculture / culture@unaki.ac.id

or highly homogeneous response patterns between the upper and lower groups. This phenomenon may occur when items are strongly aligned with dominant content areas but do not fully represent broader constructs (Tavakol & Dennick, 2023). Therefore, slight variation and diversification among items may help maintain reliability while improving the depth of assessment.

Despite the strong overall performance of the instrument, the distractor analysis revealed that several distractors were non-functional, meaning they were rarely selected or did not attract low-ability students. As emphasized by Haladyna (2004) and Nitko and Brookhart (2011), non-functioning distractors weaken item quality and can reduce test precision. Some distractors that were never chosen indicate that answer options are too obviously incorrect or that students may be familiar with the content. This suggests the need for improved distractor development, especially for items with moderate or low discrimination.

The findings demonstrate that the psycholinguistics test instrument is valid, reliable, and effective for evaluating learning outcomes. However, minor revisions—especially for items showing low discrimination or ineffective distractors—are essential for enhancing measurement accuracy and supporting continuous improvement. As supported by recent research (Hrich et al., 2024; Hartati & Yogi, 2019), software-based item analysis tools such as ITEMAN are valuable for guiding empirical decision-making in assessment and strengthening evidence-based test development in higher education settings.

5. Conclusions

Based on the results of the item analysis using ITEMAN 4.0 and reviewed from the perspective of Classical Test Theory (CTT), it can be concluded that the analyzed Psycholinguistics test instrument has excellent quality and is suitable for use as a learning evaluation tool. The majority of the items were at a moderate difficulty level, indicating that the test measured participants' competency proportionally. The high distinguishing power of most items indicates that they effectively distinguish between high- and low-ability students, in accordance with the basic principles of CTT.

The very high grain-total correlation corroborates the instrument's construct consistency and further supports its very high internal reliability, as reflected in the KR-20 value. These findings suggest that the instrument has strong homogeneity and near-perfect internal consistency. In addition, most tricks work effectively, though some need revision because the participants did not choose them or did not work as intended.

This Psycholinguistics test has met the criteria for instrument quality according to theory, both in terms of difficulty, differentiation, correlation, deception function, and reliability. For sustainability advice, minor revisions to certain items, particularly overly difficult items and malfunctioning tricks, can yield a valid, reliable, and representative final question bank to measure a student's Psycholinguistics competence.

6. References

Alderson, J. C. (2020). *Language test construction and evaluation*. Routledge.

Aljabr, A., Shaikh, S., Kannan, S. K., Aldhuwayhi, S., Jayakumar, S., Al-Roomy, R., Uthappa, R., & Alkhujairi, A. (2021). Replacing non-functional distractors to improve the quality of MCQs: A quasi-experimental study. *International Journal of Educational Sciences, 33*(1–3), 52–58. https://doi.org/10.31901/24566322.2021/33.1-3.1186

Allen, M. J., & Yen, W. M. (2022). *Introduction to measurement theory*. Waveland Press.

Ayanwale, M. A. (2022). Classical test theory and item response theory: A comparative review. *Journal of Measurement and Evaluation, 21*(8). https://doi.org/10.26803/ijlter.21.8.22

e-ISSN 2775-4618, p-ISSN 2355-8660, bit.ly/jurnalculture / culture@unaki.ac.id

Azwar, S. (2017). *Reliability and validity*. Student Library.

Brown, H. D., & Abeywickrama, P. (2021). *Language assessment: Principles and classroom practices* (4th ed.). Pearson Education.

Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. Cengage Learning.

DeVellis, R. F., & Thorpe, C. T. (2021). *Scale development: Theory and applications* (5th ed.). SAGE.

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement*. Prentice Hall.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items*. Lawrence Erlbaum Associates.

Hambleton, R. K., & Jones, R. W. (2021). *Modern measurement theory*. Routledge.

Hartati, N., & Yogi, H. P. S. (2019). Item analysis for better quality test. *English Language in Focus (ELIF), 2*(1), 59–70. https://jurnal.umj.ac.id/index.php/ELIF

Hrich, N., Azekri, M., & Khaldi, M. (2024). Artificial intelligence item analysis tool for educational assessment: Case of large scale competitive exams. *International Journal of Information and Educational Technology, 14*(4), 822–827. https://www.ijiet.org/vol14/IJIET-V14N6-2107.pdf

Kline, P. (2020). *Psychological testing: A practical approach to design and evaluation*. Routledge.

Liu, Y., Pek, J., & Maydeu-Olivares, A. (2024). *Understanding reliability from a regression perspective*. https://doi.org/10.48550/arXiv.2404.16709

Miller, L. A., & Lovler, R. L. (2020). *Foundations of psychological testing: A practical approach* (6th ed.). Sage Publications.

Nitko, A. J., & Brookhart, S. M. (2011). *Educational assessment of students*. Pearson.

Nurjanah, S., Rahmawati, D., & Wicaksono, A. (2024). Psychometric quality of multiple-choice tests under Classical Test Theory using ITEMAN software. *Jurnal Evaluasi Pendidikan, 15*(2), 101–118. https://jurnal.uny.ac.id/index.php/jpep/article/download/71542/23430

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.

Nuroini, N., & Syamsudin, A. (2023). Item quality analysis of final semester test in Chemistry subject using ANATES. *Jurnal Kimia dan Pendidikan Kimia, 7*(1), 55–66. https://jurnal.uns.ac.id/jkpk/article/view/54999

Ohiri, S. C., & Okoye, R. O. (2024). Application of classical test theory as linear modeling to test item development and analysis. *International Research Journal of Modernization in Engineering, Technology and Science, 5*(10). https://doi.org/10.56726/IRJMETS45379

Resi, D. N. (2025). Quality analysis of Biology mid-semester assessment questions with classical test theory and Rasch model. *Bio-Pedagogy, 13*(2), 63. https://doi.org/10.20961/bio-pedagogi.v13i2.88075

Rezigalla, A., Abdalla, O., & Ibrahim, S. (2024). Item analysis: The impact of distractor efficiency on the difficulty index and discrimination power of multiple-choice items. *Research in Medical Education, 13*(1), 22–29.

Rohmatdi, A. (2024). Classical test theory evaluation with ITEMAN 4.3. *Al-Ishlah: Journal of Education, 16*(4). https://journal.staihubbulwathan.id/index.php/alishlah/article/download/5671/2568

Shah, R., Patel, S., & George, A. (2019). Item analysis as a tool to validate multiple-choice questions. *International Journal of Basic & Clinical Pharmacology, 8*(3), 540–545. https://www.ijbcp.com/index.php/ijbcp/article/view/3324

e-ISSN 2775-4618, p-ISSN 2355-8660, bit.ly/jurnalculture / culture@unaki.ac.id

Subhaktiyasa, P. G. (2024). Evaluation of the validity and reliability of quantitative research instruments: A literature study. *Journal of Education Research, 5*(4), 5599–5609. https://doi.org/10.37985/jer.v5i4.1747

Tavakol, M., & Dennick, R. (2023). Multiple-choice item analysis: A contemporary review. *Medical Education Review, 57*(2), 215–230.

Thompson, N. A. (2022). Classical test theory in modern assessment practice. *Educational Measurement Review, 18*(1), 25–38.

Traxler, M. J. (2023). *Introduction to psycholinguistics: Understanding language science* (2nd ed.). Wiley-Blackwell.

Zubairi, N. A., Al-Haqan, A., Al-Fadhli, S., & Al-Mutairi, R. (2025). Effective use of item analysis to improve the reliability and validity of multiple-choice examinations. *Journal of Taibah University Medical Sciences, 20*(1), 1–10. https://pmc.ncbi.nlm.nih.gov/articles/PMC11911747/